

Galaxy

Workshop spring school 2019

The screenshot displays the Galaxy@IPK web interface. The top navigation bar includes links for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Admin', 'Help', and 'User'. The main content area is titled 'Galaxy@IPK' and features three primary sections: 'First steps' (with a gear icon), 'Tutorials' (with a lightning bolt icon), and 'You need help?' (with a star icon). The left sidebar lists various tool categories such as 'BASIC TOOLS', 'NCBI BLAST+', and 'NGS DATA MANIPULATION'. The right sidebar shows a 'History' panel with a list of datasets, including 'intermedia' and '620: Pasted Entry'.

A short introduction to Galaxy

- Introduction of Galaxy's main functionality
 - Creating a login for usegalaxy.eu

https://usegalaxy.eu/join-training/ipk_springschool
- Coffee break
- Hands-on: First steps on usegalaxy.eu
 - Uploading data
 - Running tools
 - Creating a simple workflow
 - Setting up a dataset collection
 - (Run the RNA-Seq workflow)

What is Galaxy?

“Galaxy is an open, web-based platform for accessible, reproducible and transparent computational research.”

<http://galaxyproject.org>

- Developed at Penn State, Johns Hopkins, OHSU and Cleveland Clinic with substantial outside contributions
- **Open source** under Academic Free License
 - 116 public servers (both general-purpose and domain-specific)
 - Numerous internal servers at institutes and companies
 - Citizen science projects

Galaxy-E initiative (<https://github.com/65MO/Galaxy-E>)

- Initially devoted to citizen science projects related to Biodiversity started in 2015
- When uploading a dataset, its datatype can be project coordinated of the french National Museum of Natural History MNHN called "65 Millions d'observateurs"
 - Bird observation
- Users are directly enrolled in the analysis

Members of the public are invited to take part in a national bird counting survey in their gardens, on January 27 and 28.

For the sixth year running, the biodiversity network the LPO and national museum the *Muséum national d'Histoire naturelle (MNHN)* are joining together in the initiative, which seeks to follow bird populations living in proximity with humans, in a bid to understand their condition and to help develop measures to protect the animals.

Everyone can take part, whether from a city or the countryside, a big garden or small yard, or even from a balcony or windowsill. Even a public park counts.

Only a few words of French (to understand the website when submitting your details) and a connection to the internet are needed to add your observations - there is no need to be a bird or nature expert.



Oldiefan / CC0 Creative Commons // Gerrit Vyn / The Cornell Lab of Ornithology // Dfaulder / CC BY 2.0

...as a Bioinformatician?

Open Source, python-based downloadable package that can be deployed in individual labs:

- modularized
- add new tools (maintain various versions)
- integrate new data sources
- easely plug in your own components

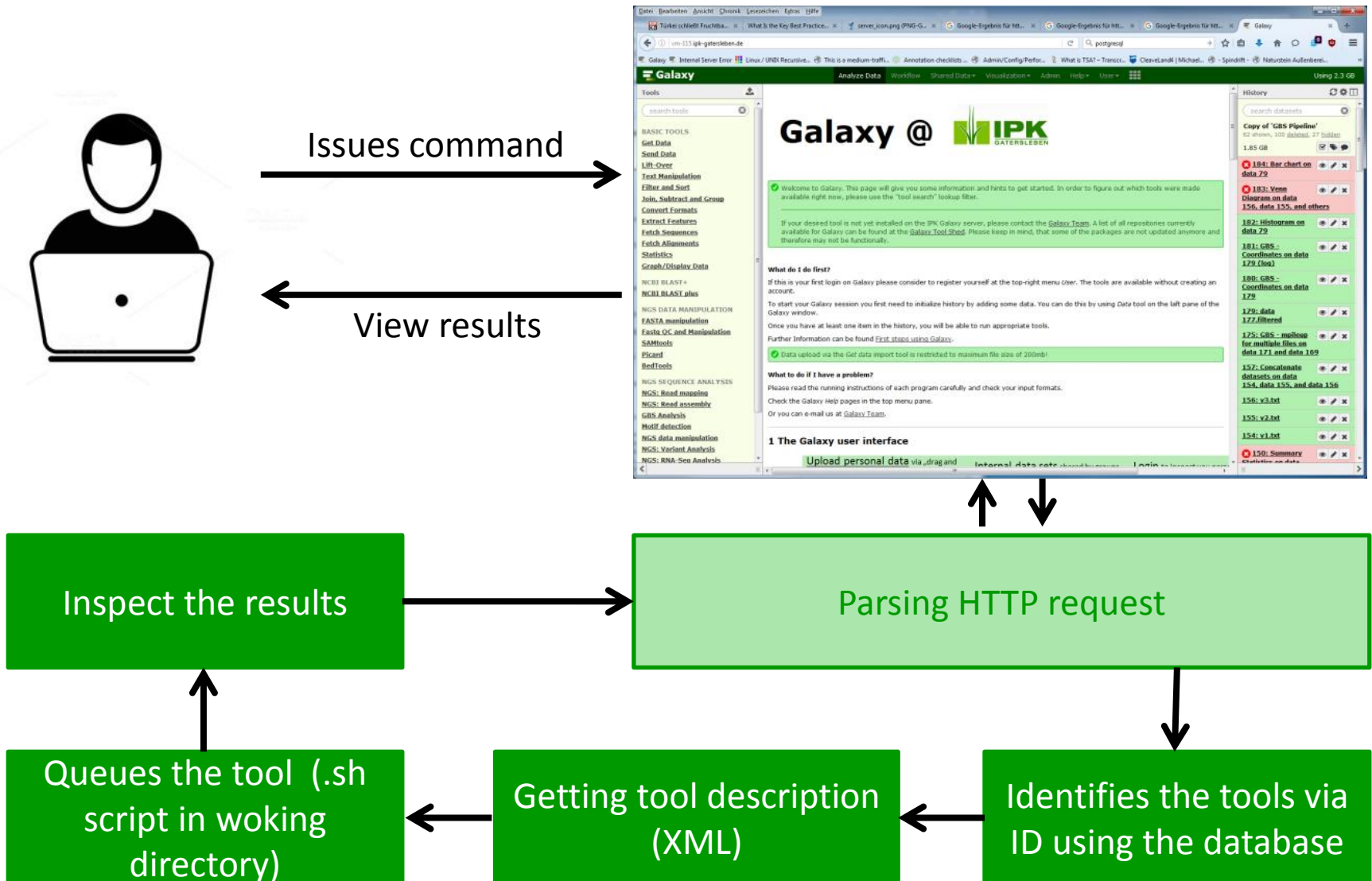
...as a scientist without bioinformatic background?

You can:

- analyze genome-scale NGS data without bash scripting
- work with big datasets, genomic regions, sequences etc.
- create and use Galaxy workflows
- share results and workflows with a user or make it available to anyone

1 Overview

Galaxy Workflow



1

Overview

The IPK Galaxy user interface

The screenshot shows the Galaxy web interface with several key areas highlighted and annotated:

- Top Navigation Bar:**
 - Analyze Data:** Circled in red, with an arrow pointing to the text "„Home“ go back to the homepage".
 - Workflow:** Circled in green, with an arrow pointing to the text "Workflows create or edit pipelines".
 - Shared Data:** Circled in green, with an arrow pointing to the text "Internal datasets and workflows shared by users/groups".
 - User:** Circled in green, with an arrow pointing to the text "User data to manage you personal datasets/histories."
- Left Panel (Tools):**
 - A green arrow points from the **Get Data** link to the text "Upload personal data via „drag and drop“ directly from you computer (<2GB file size)".
 - The **Get Data** link is circled in green.
 - Other tool categories listed include: Send Data, Lift-Over, Text Manipulation, Filter and Sort, Join, Subtract and Group, Fasta Manipulation, Convert Formats, Extract Features, Fetch Sequences, Fetch Alignments, Statistics, Graph/Display Data, NCBI BLAST plus, Mapping, SAMtools, Picard, Motif detection, NGS data manipulation, Genome assembly, and Workflows.
- Working Desktop (Center):**
 - Contains the **Minimus2** tool interface for merging contig sequences.
 - Fields include: "Contig sequences file 1" (58: CS1.Trinity.fasta), "Rename contigs in file 1 by adding prefix?" (yes, add prefix), "Prefix for sequences in file 1" (CS1), "Contig sequences file 2" (59: CS2.Trinity.fasta), "Rename contigs in file 2 by adding prefix?" (yes, add prefix), and "Prefix for sequences in file 2" (CS2).
 - An **Execute** button is at the bottom.
 - Below the form is a "What it does" section explaining that Minimus2 is part of the AMOS assembler package and preprocesses Fasta input files.
 - A "Documentation" link is provided at the bottom.
- Right Panel (History):**
 - Contains a list of datasets and jobs.
 - Items include: "513.3 MB", "59: CS2.Trinity.fasta", "58: CS1.Trinity.fasta", "50: test.fa", "43: IdxStats on data 41", "42: Flagstat on data 41", "41: Map with BWA-MEM on data 40, data 39, and data 38 (mapped reads in BAM format)", "40: LS2 R2.achtel.trim.fg", "39: LS2 R1.achtel.trim.fg", and "38: Brassica napus v4.1.chromosome s.fa".
 - Each item has icons for viewing, editing, and deleting.

Bottom Summary:

- Available analysis tools:** Click on folders to open subcategories
- Working Desktop**
- History:** Find your uploaded data and jobs here, clicking the name results in extended information on the job

The screenshot displays the Galaxy web interface. On the left, the 'Tools' sidebar is visible, with the 'merge' tool highlighted in a red box. A green arrow points from this box to the main tool configuration area. The main area shows the 'Merge Columns together (Galaxy Version 1.0.1)' tool. The 'Select data' section has a dropdown menu showing '616: megablast Pasted Entry vs duckweed_9509'. The 'Merge column' section has a dropdown menu showing 'Column: 1'. The 'with column' section also has a dropdown menu showing 'Column: 1'. There is a green 'Execute' button at the bottom of the configuration area. A tip below the button states: 'TIP: If your data is not TAB delimited, use Text Manipulation->Convert'. The right sidebar shows the 'History' tab with a search bar and a list of datasets, including '568: megablast polyrhiza 5S 2 vs duckweed.paper', '567: polyrhiza 5S 2', and '566: megablast polyrhiza 5S vs'.

- Organised in categories
- The tool search helps in finding a tool in a crowded toolbox

HISAT A fast and sensitive alignment program (Galaxy Version 2.0.3) [Versions](#) [Options](#)

Input data format

FASTQ

Single end or paired reads?

Collection of paired reads

Paired reads

No fastq dataset collection available.

Paired-end options

Use default values

Source for the reference genome to align against

Use a built-in genome

Built-in references were created using default options

Select a reference genome

Aegilops tauschii AL8 (UC Davis v.78,Nov 2017)

If your genome of interest is not listed, contact the Galaxy team

Primary alignments

5

Search for at most K distinct, primary alignments for each read. Primary alignments mean alignments whose alignment score is equal or higher than any other alignments. The search terminates when it can't find more distinct valid alignments, or when it finds K, whichever happens first. (-k)

Alignment options

Use default values

Input options

Use default values



Scoring options

Use default values

Spliced alignment parameters

Use default values

- Tools are simply text files with:
 - input datasets, parameters, commands, and outputs
 - help, tests, citations, dependency requirements
- New versions can be installed without removing old ones to ensure reproducibility

HISAT A fast and sensitive alignment program (Galaxy Version 2.0.3)  Versions  Options

Input data format
FASTQ

Single end or paired reads?
Collection of paired reads

Paired reads
No fastq dataset collection available.

Paired-end options
Use default values

Source for the reference genome to align against
Use a built-in genome
Built-in references were created using default options
Select a reference genome
Aegilops tauschii AL8 (UC Davis v.78,Nov 2017)
If your genome of interest is not listed, contact the Galaxy team

Primary alignments
5
Search for at most K distinct, primary alignments for each read. Primary alignments mean alignments whose alignment score is equal or higher than any other alignments. The search terminates when it can't find more distinct valid alignments, or when it finds K, whichever happens first. (-k)

Alignment options
Use default values

Input options
Use default values

Scoring options
Use default values

Spliced alignment parameters
Use default values

HISAT A fast and sensitive alignment program (Galaxy Version 2.0.3) [Versions](#) [Options](#)

Input data format

FASTQ

Single end or paired reads?

Collection of paired reads

Paired reads

No fastq dataset collection available.

Paired-end options

Use default values

Source for the reference genome to align against

Use a built-in genome

Built-in references were created using default options

Select a reference genome

Aegilops tauschii AL8 (UC Davis v.78,Nov 2017)

If your genome of interest is not listed, contact the curator

Primary alignments

5

Search for at most K distinct, primary alignments for each read. The search terminates when a higher than any other alignments. The search terminates when the first alignment happens first. (-k)

Alignment options

Use default values

Input options

Use default values

Scoring options

Use default values

Spliced alignment parameters

Use default values

Input options

Specify input parameters

Skip the first N reads or pairs in the input

0

(-s)

Stop after aligning N reads

0

Align the first N reads or read pairs from the input (after the first N reads or pairs have been skipped), then stop. (-u)

Trim 5' end

0

Trim N bases from 5' (left) end of each read before alignment (-5)

Trim 3' end

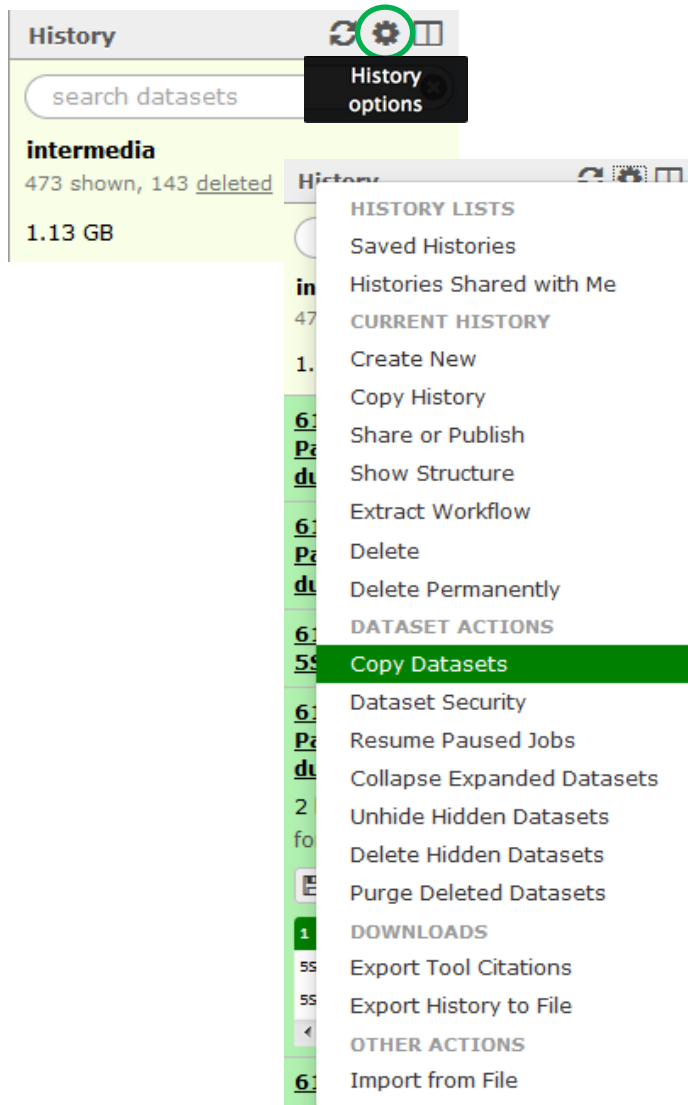
0

Trim N bases from 3' (right) end of each read before alignment (-3)

Scoring options

- Location of all analyses
 - Collects all datasets produced by tools
 - Collects all operations performed on the data
- For each dataset (the heart of Galaxy's reproducibility), the history tracks
 - Name, format, size, creation time, datatype-specific metadata
 - Tool ID, version, inputs, parameters
 - Standard output (stdout) and error (stderr)
 - State (waiting, running, success, failed)
 - Hidden, deleted, purged

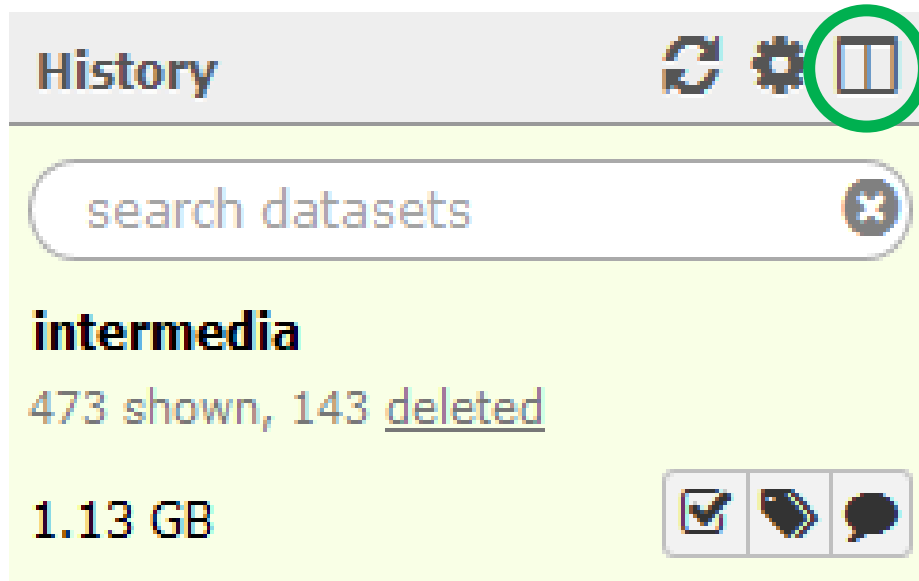
VPE 166 shown, 93 deleted 7.72 MB			<input checked="" type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>
<input type="text" value="search datasets"/>			<input type="button" value="x"/>
247: megablast VPE_orf vs wheat_nr.pseudo	<input type="button" value="eye"/>	<input type="button" value="pencil"/>	<input type="button" value="x"/>
<input checked="" type="checkbox"/> 246: megablast VPE_orf vs wheat_nr.pseudo	<input type="button" value="eye"/>	<input type="button" value="pencil"/>	<input type="button" value="x"/>
<input checked="" type="checkbox"/> 245: megablast VPE_orf vs wheat_nr.pseudo	<input type="button" value="eye"/>	<input type="button" value="pencil"/>	<input type="button" value="x"/>
244: Extract Genomic DNA on data 243	<input type="button" value="eye"/>	<input type="button" value="pencil"/>	<input type="button" value="x"/>
243: weizen_NR.bed	<input type="button" value="eye"/>	<input type="button" value="pencil"/>	<input type="button" value="x"/>
242: megablast VPE_orf vs wheat_nr.pseudo	<input type="button" value="eye"/>	<input type="button" value="pencil"/>	<input type="button" value="x"/>
<input checked="" type="checkbox"/> 241: megablast VPE_orf vs wheat_nr.pseudo	<input type="button" value="eye"/>	<input type="button" value="pencil"/>	<input type="button" value="x"/>
240: tblastn VPEd protein vs wheat_nr4	<input type="button" value="eye"/>	<input type="button" value="pencil"/>	<input type="button" value="x"/>
239: tblastn VPEc protein vs wheat_nr4	<input type="button" value="eye"/>	<input type="button" value="pencil"/>	<input type="button" value="x"/>
238: tblastn VPEb protein vs wheat_nr4	<input type="button" value="eye"/>	<input type="button" value="pencil"/>	<input type="button" value="x"/>
237: tblastn VPEa protein vs wheat_nr4	<input type="button" value="eye"/>	<input type="button" value="pencil"/>	<input type="button" value="x"/>
<input checked="" type="checkbox"/> 227: JBrowse on - Complete	<input type="button" value="eye"/>	<input type="button" value="pencil"/>	<input type="button" value="x"/>
226: Extract features on data 225	<input type="button" value="eye"/>	<input type="button" value="pencil"/>	<input type="button" value="x"/>
225: affread on : aff3	<input type="button" value="eye"/>	<input type="button" value="pencil"/>	<input type="button" value="x"/>



History behavior is controlled by the *History options* (gear icon)

- *Create New* history will **not** make your current history disappear
- *Copy Datasets* from one history to another and save disk space for your quota
- Several delete options

- You can have as many histories as you want
 - Each history should correspond to a **different analysis**
 - Should have a meaningful **name**
- To see all of your histories, use the history switcher



- You can have as many histories as you want
 - each history should correspond to a **different analysis**
 - and should have a meaningful **name**

The screenshot displays the Galaxy web interface with five histories visible in a row. Each history has a title, a description, a size, and a list of datasets. The 'Current History' is highlighted in green.

Galaxy Analyze Data Workflow Shared Data Visualization Admin Help User Using 4.8 GB

Done search histories search all datasets Create new

Current History Switch to

intermedia
473 shown, 143 deleted
1.13 GB
search datasets
Drag datasets here to copy them to the current history

616: megablast Pasted Entry vs duckweed 9509
615: megablast Pasted Entry vs duckweed 7498
614: 5S_polyrhiza.fa
613: megablast Pasted Entry vs duckweed v2
2 lines
format: tabular, database: ?

1	2	3	4	5	6	7	8	9	10
55_cons_pseudo6	97.84	139	3	0	1	139	3523797	3523659	
55_cons_pseudo8	99.16	119	1	0	1	119	1598697	1598579	

612: Pasted Entry
611: megablast Pasted Entry vs intermedia_scaf.4
610: Pasted Entry
609: megablast Pasted Entry vs intermedia_scaf.4
72 lines
format: tabular, database: ?

1 2 3 4 5 6 7 8

RNA-Seq for two conditions Switch to

6 shown, 73 hidden
54.4 MB
search datasets

79: All samples TPM
78: All raw counts
34,725 lines
format: tabular, database: tomato.chr

1	2	3	4	5
KEY	1.fastq_0	2.fastq_0	3.fastq_6	4.fe
mRNA:Solyc08g005020.1.1	0	0	0	0
mRNA:Solyc08g005040.2.1	0	0	0	0
mRNA:Solyc08g005050.2.1	2	4	2	2
mRNA:Solyc08g005060.1.1	0	0	0	0

76: All featureCounts statistics
75: DESeq2 plots on data 69, data 68, and others
74: DESeq2 result file on data 69, data 68, and others
29: HISAT2 on control
a list with 3 items

Test Switch to

159 shown, 96 deleted, 344 hidden
251.79 MB
search datasets

501: featureCounts on data 484
500: All raw counts for test_neuehisat
499: FeatureCounts statistics for test_neuehisat
11 lines
format: tabular, database: tomato.chr

1	2	3
Assigned	4.fastq_33629	5.fastq_33572
Unassigned_Ambiguity	533	511
Unassigned_Chimera	0	0
Unassigned_Duplicate	0	0
Unassigned_FragmentLength	0	0

488: featureCounts on collection
483: summary
a list with 3 items
487: FeatureCounts on test_neuehisat
a list with 3 items
468: DESeq2 plots on data 459, data 457, and others
467: DESeq2 result file on data 459, data 457, and others
466: All raw counts for

Unnamed history Switch to

24 shown
(empty)
search datasets

24: 6 GTT L0 R2 002.fastq.qz
23: 6 GTT L0 R2 001.fastq.qz
22: 6 GTT L0 R1 002.fastq.qz
21: 6 GTT L0 R1 001.fastq.qz
20: 5 GTT L0 R2 002.fastq.qz
19: 5 GTT L0 R2 001.fastq.qz
18: 5 GTT L0 R1 002.fastq.qz
17: 5 GTT L0 R1 001.fastq.qz
16: 4 GTT L0 R2 002.fastq.qz
15: 4 GTT L0 R2 001.fastq.qz
14: 4 GTT L0 R1 002.fastq.qz
13: 4 GTT L0 R1 001.fastq.qz
12: 3 GTT L0 R2 002.fastq.qz

Tomato RNA-Seq edgeR Switch to

131 shown, 35 deleted, hide hidden
104.58 MB
search datasets

This dataset has been hidden
Unhide it
166: Trimmomatic on 3 2.fastq (R2 unpaired)
This dataset has been hidden
Unhide it
165: Trimmomatic on 3 1.fastq (R1 unpaired)
This dataset has been hidden
Unhide it
164: Trimmomatic on 3 2.fastq (R2 paired)
This dataset has been hidden
Unhide it
163: Trimmomatic on 3 1.fastq (R1 paired)
This dataset has been hidden
Unhide it
162: Trimmomatic on 2 2.fastq (R2 unpaired)
This dataset has been hidden
Unhide it
161: Trimmomatic on 2 1.fastq (R1 unpaired)
This dataset has been hidden
Unhide it

2

Data import

Data upload

Galaxy

Analyze Data Workflow Shared Data Visualization Admin Help User

Tools

search tools

BASIC TOOLS

Get Data

Send Data

Lift-Over

Text Manipulation

Filter and Sort

Join, Subtract and Group

Convert Formats

Extract Features

Fetch Sequences

Fetch Alignments

Statistics

Graph/Display Data

NCBI BLAST+

NCBI BLAST plus

NGS SEQUENCE ANALYSIS

FASTA manipulation

NGS: Read assembly

NGS: Read mapping

GBS Analysis

SAMtools

Motif detection

Picard

NGS data manipulation

NGS: Variant Analysis

Download data directly from web or upload files from your disk

You can Drag & Drop files into this box.

You added 1 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Type	Genome	Settings	Status
input	1.2 KB	Auto-det...	unspecified (?)		

Type (set all): Auto-detect Q Genome (set all): unspecified (?)

Choose local file Paste/Fetch data Start Pause Reset Close

History

search datasets

test

10 shown, 54 deleted

442.77 MB

64: FASTQ joiner on data 62

442.8 MB

format: fastqsanger, database: ?

Joined 1000000 of 1000000 read pairs (100.00%).

62: TrSc4B_MIO_R2.fastq

52: Adapter

1 sequences

format: fasta, database: ?

uploaded fasta file



Input data file

Go to **Get data** > **Upload file** from your computer:

A) Move your data file (<200mb) from your computers desktop into the white popup.

or

B) Choose the location of your files (preferred if your have multiple files to upload)

or

C) Paste data directly using „Copy/Paste“ operations (keyboard, mouse)

Press the **Start** button to complete the import

Data libraries are a convenient framework within Galaxy to store and share data.

Data can be shared with you:


- by the Galaxy admins (e.g. sequencing data from HSM)
- by another user

The screenshot shows the Galaxy Data Libraries interface. At the top, there's a navigation bar with 'Workflow', 'Shared Data', and 'Visualization'. A dropdown menu for 'Shared Data' is open, showing 'Data Libraries', 'Histories', 'Workflows', 'Visualizations', and 'Pages'. Below this, the 'DATA LIBRARIES' section shows a list of libraries. The first library is 'GFF/GTF data' with the description 'Annotation files for reference fastas'. The second library is 'HSM' with the description 'HSM Sequenzen'. A green arrow points from the 'HSM' library to the detailed view below.

The detailed view shows the 'HSM' library with 12 items. The table below lists the items:

name	description	data type	size	time updated (UTC)
Lo115_CGATGT_L003_R1_001.fastq		fastq	15.8 GB	2016-08-15 11:24 AM
Lo115_CGATGT_L003_R2_001.fastq		fastq	15.7 GB	2016-08-15 11:24 AM
Lo115_CGATGT_L004_R1_001.fastq		fastq	15.9 GB	2016-08-15 11:24 AM
Lo115_CGATGT_L004_R2_001.fastq		fastq	15.9 GB	2016-08-15 11:24 AM

A green arrow points from the 'to History' button in the top right of the detailed view to the 'Your history' text on the left. A large green arrow on the left points from the 'Your history' text to the detailed view.

- Tools only accept input datasets with the appropriate datatypes!
- When uploading a dataset, its datatype can be either:
 - automatically detected (fasta, txt, tabular)
 - assigned by user (fastq, other datatypes)
- Dataset produced by a tool: datatype assigned by the tool
- To change the datatype of a dataset in history:
 - *Edit Attributes and Datatype*
 - *Edit Attributes and Convert Formats* 

- Sequences (FASTA, FASTQ, ABI/SCF, SFF)
- Alignments (MAF, SAM/BAM, AXT, LAV)
- Intervals (BED, INTERVAL, GFF(3), WIG)
- Tabular data (e.g. CSV)
- Others (HTML, TXT, LPED)
- Compressed file formats accepted (.gz, .zip)
- Format conversion
 - Converters included
- Not supported
 - MS Office binaries (Excel, Word) => export as „txt“ or „csv“

- Private data
 - Upload from your own system (or copy/paste)
 - Import from shared data library (provided by the admins to keep restricted datasets confidential)
 - Private BLAST databases can included (also with restriction to special users/groups at IPK)
- Public data
 - [Reference genomes from web services (*UCSC Genome Browser*)]
 - From URL (e.g. import from ENA SRA or NCBI SRA)



Navigation Read Files

This table contains the files for run SRR2240229
[Download files](#)

Download: - of 1 results in [TEXT](#)

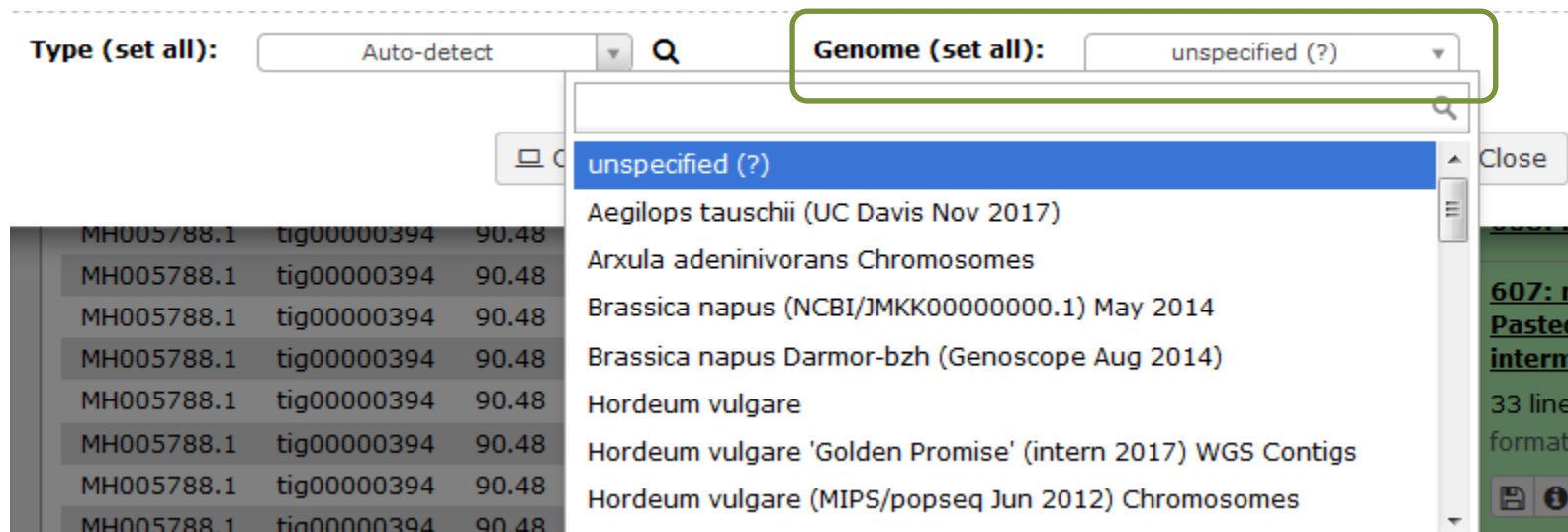
[Select columns](#)

Showing results 1 - 1 of 1 results

Study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Tax ID	Scientific name	Instrument model	Library layout	Fastq files (ftp)	Fastq files (galaxy)	Submitted files (ftp)	Submitted files (galaxy)	NCBI SRA file (ftp)	NCBI SRA file (galaxy)	CRAM Index files (ftp)	CRAM Index files (galaxy)
PRJNA293270	SAMN04027991	SRS1053908	SRX1181347	SRR2240229	4513	Hordeum vulgare	Illumina HiSeq 2000	SINGLE	File 1	File 1			File 1	File 1		

Reference genomes

- Genome build specifies which genome assembly a dataset is associated with
- Can be automatically detected or assigned by user
- New builds can be added by the admin
- Some tools allow to create indices from uploaded files (blast DBs)
 - Con: very time consuming for whole genome files (plants)



3

Running Galaxy

Processing your job

1

After submitting the job, a new box will be created, showing the job progress:

Job submitted

Job running

Job completed or Job failed

The color code indicates the job progress

2

Your history panel with the uploaded dataset „input“ (B)

History

search datasets

Unnamed history
25 shown, 4 deleted

48.1 KB

29: blastp input vs SorghumBiColor GeneModels S bi1 4 aa2.fasta

26: input

3

View your data

Click to open

Inspect and edit attributes

Edit Attributes

Name:
input

Info:
uploaded fasta file

Annotation / Notes:
Add an annotation or notes to a dataset; annotations are available when a history is viewed.

Database/Build:
unspecified (?)

Save

Auto-detect

This will inspect the dataset and attempt to correct the above column values if they are not accurate.

Tool: Upload File

Name: input

Created: Fri Dec 18 09:48:08 2015 (UTC)

Filesize: 1.2 KB

Dbkey: ?

Format: fasta

Galaxy Tool ID: upload1

Galaxy Tool Version: 1.1.4

Tool Version:

Tool Standard: standard

Output: standard

Error:

Tool Exit Code: 0

API ID: 9a46a440b852c3ba

History ID: 5a93819f2a7b69af

UUID: 9a46a440b852c3ba-4425-9499-3132109ec035

Full Path: /opt/galaxy/database/files/003/dataset_3257.dat

Job: python /opt/galaxy/tools/data_source/upload.py /opt/galaxy /opt/galaxy/database/tmp/tmpICUFNa /opt/galaxy/database/tmp/tmpR1X8KH

Command Line: 3257 /opt/galaxy/database/job_working_directory/000/416/dataset_3257_files /opt/galaxy/database/files/003/dataset_3257.dat

Job End Time: 2015-12-18 10:48:19

Job Start Time: 2015-12-18 10:48:13

Cores Allocated: 1

Job Runtime (Wall Clock): 6 seconds

Details how the file was generated

Input Parameter	Value	Note for rerun
File Format	auto	
async_datasets	None	
Specify Files for Dataset (auto)	1 uploaded datasets	
Genome	unspecified (?)	
File Format	auto	

Download the file to your workspace

Re-run the job using the same configuration (or change some parameters)

25: input

2 sequences

format: fasta, database: ?

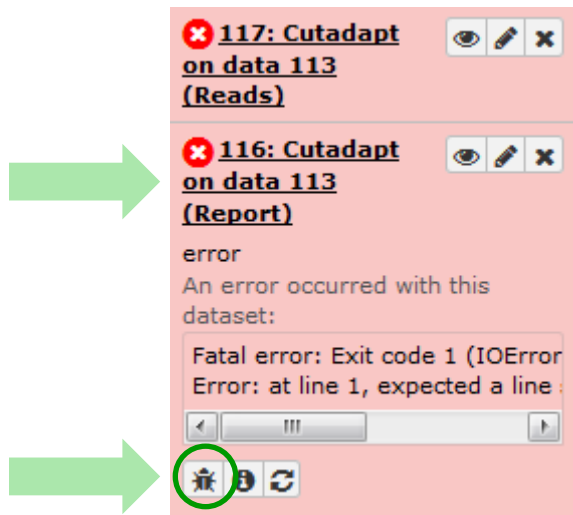
uploaded fasta file

Preview of the data

```
>GSERNA2T0000001001 assembled CDS
MATGENRTVQENLKKHLAVSVRNQISVGIFWISISASQPGVL
AVEVKADQLGLERSDQLRELYESLSVAESSASGGGSQVSRAS
CNSFVFNIGEGITGGALNGEPIWLQVHTADSKVFTRSLLA
EIGTTEHTEDFNWICVKTLFLEAHPYGTISTRSDVQEIFD
TSTSVEFEQLEDHDSFINGGSASQVQSQFVGEELNNCVHQP
```

Delete your data

- Check your input data
- Read the error report!



- Search for the error message
- <https://biostar.usegalaxy.org>
- Report the error to the local Galaxy administrators

Dataset generation errors

Dataset 116: Cutadapt on data 113 (Report)

Tool execution generated the following error message:

```
Fatal error: Exit code 1 (IOError, FormatError, or Interrupt)
Error: at line 1, expected a line starting with '+'
```

Report this error to the local Galaxy administrators

Usually the local Galaxy administrators regularly review errors that occur on the server. However, if you would like to provide additional information (such as what you were trying to do when the error occurred) and a contact e-mail address, we will be better able to investigate your problem and get back to you.

Error Report

Your email

Message

Report



A workflow is ..

.. a series of tools and dataset actions that run in sequence as a batch operation.

- Workflows specify the steps in a process.
- Workflows are analysis that are meant to be run, each time with different user provided datasets.
- Reproducible and well documented bioinformatic pipelines

Running workflow "Copy of 'GBS Pipeline'"

Primary analysis of genotyping-by-sequencing (GBS) data of barley

[Expand All](#)
[Collapse](#)

Step 1: Input dataset collection

Input Dataset Collection

6: Map with BWA-MEM on collection 117 (mapped reads in BAM for
type to filter

Step 2: Map with BWA-MEM (version 0.7.12.1)

Step 3: Novosort (version alpha)

Step 4: GBS - mpileup for multiple files (version 1.0.0)

Step 5: bcftools call (version 1.0)

Step 6: GBS - gen call in python (version 1.0.0)

VCF file to filter

Output dataset 'vcf_out' from step 5

Minimum SNP quality

40 

Minimum quality for a homozygous genotype call

3 

Minimum quality for a heterozygous genotype call

5 

Minimum read depth for a homozygous genotype call

1 

Minimum read depth for a heterozygous genotype call

3 

Maximum fraction of missing genotype calls

0.9 

Minimum minor allele frequency

0.05 

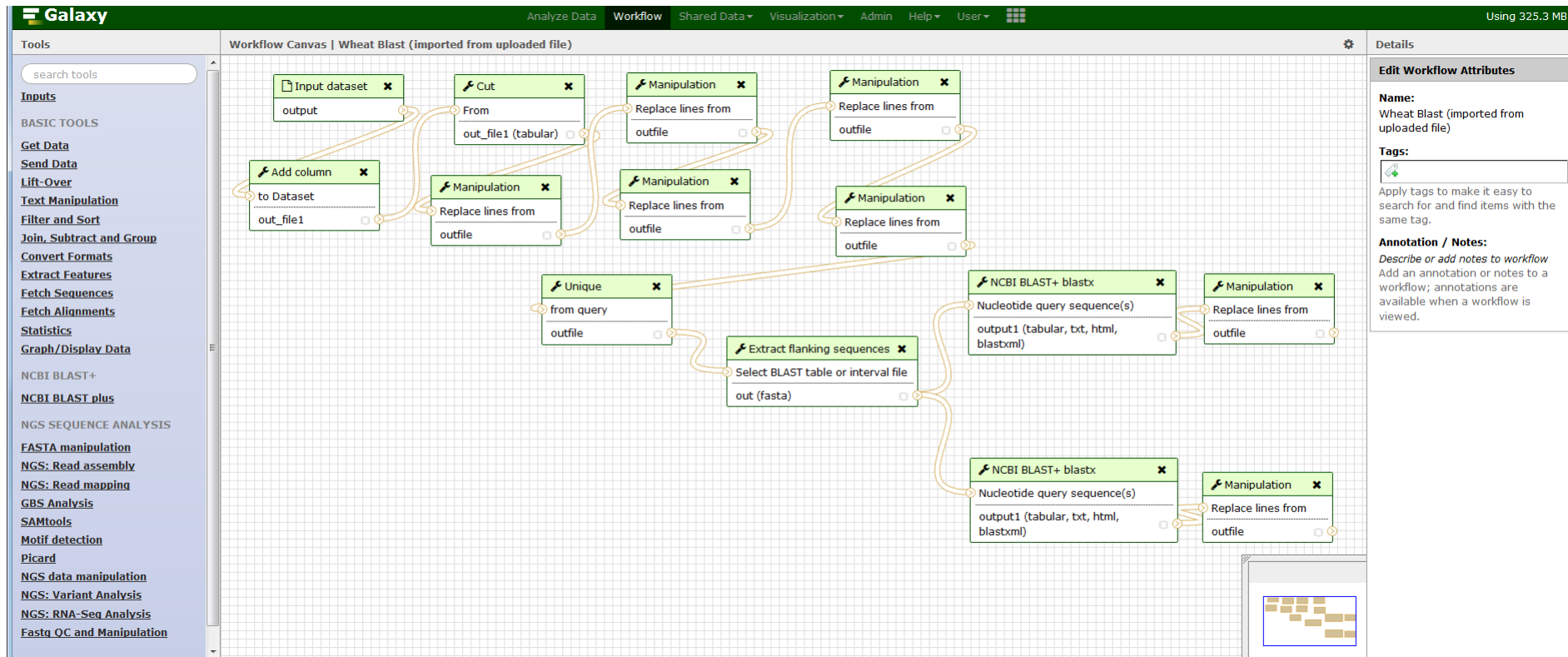
Minimum fraction of heterozygous calls

0.1 

Step 7: GBS - Coordinates (version 1.0.0)

☐ Send results to a new history

Run workflow



Workflows can be created from scratch using the *workflow editor*

Or...

The screenshot displays the Galaxy workflow editor interface. The main canvas shows a workflow titled "Wheat Blast (imported from uploaded file)". The workflow consists of several tools connected by arrows, representing a data processing pipeline. The tools include "Input dataset", "Cut", "Manipulation", "Add column", "Unique", "Extract flanking sequences", and "NCBI BLAST+ blastx". The left sidebar lists various tool categories such as "Inputs", "BASIC TOOLS", "Get Data", "Text Manipulation", "Filter and Sort", "Join, Subtract and Group", "Convert Formats", "Extract Features", "Fetch Sequences", "Fetch Alignments", "Statistics", "Graph/Display Data", "NCBI BLAST+", "NCBI BLAST plus", "NGS SEQUENCE ANALYSIS", "FASTA manipulation", "NGS: Read assembly", "NGS: Read mapping", "GBS Analysis", "SAMtools", "Motif detection", "Picard", "NGS data manipulation", "NGS: Variant Analysis", "NGS: RNA-Seq Analysis", and "Fastq QC and Manipulation". The right sidebar shows the "History" panel, which lists various actions and datasets, including "HISTORY LISTS", "Saved Histories", "Histories Shared with Me", "HISTORY ACTIONS", "Create New", "Copy History", "Share or Publish", "Show Structure", "Extract Workflow", "Delete", "DATASET ACTIONS", "Copy Datasets", "Dataset Security", "Resume Paused Jobs", "Collapse Expanded Datasets", "Unhide Hidden Datasets", "Delete Hidden Datasets", "DOWNLOADS", "Export Tool Citations", "Export History to File", "OTHER ACTIONS", and "Import from File".

You can specify the steps in one process using the workflow editor.

or

Doing your analysis and afterwards turn the steps in your history to a workflow.

or...

Galaxy

Analyze Data Workflow Shared Data Visualization Help User Using 0%

Published Workflows

search name, annotation, owner, and tags Q

Advanced Search

Shared Data Visualization Help User

Name	Annotation	Owner	Community Rating	Community Tags	Last Updated
DEG:Tuxedo2 - 2 Conditions 3 Replicates each	Tuxedo DEG pipeline to compare 2 conditions	setempler	★★★★★		Oct 09, 2015
RNA-seq differential analysis (single-end short reads, 2 conditions, 2 replicates)	Workflow based on Tophat and cuffdiff. Inputs: 4 fastq files (experiments), 1 bam file (pseudoreads), 1 gtf file (annotations). Outputs: bam, bigwig, xls,...	rna-seq-helin-group	★★★★★	illumina rna-seq cuffdiff tophat	Jul 17, 2013
Basic RNA-Seq Analysis - Differential Expression (Functional Genomics Workshop 2012)	From the RNA-Seq analysis tutorial during the Functional Genomics Workshop 2012 https://caps.osu.edu/pfg-workshop	mejia-guerra	★★★★★		Jun 22, 2012
Bristol workflow to get sorted unique proper pair mapped reads	This experimental workflow was designed for rna seq analysis of paired end reads. It creates a set of reads that are sorted, unique and mapped in a proper...	davidmatthews	★★★★★	unique seq rna paired sam	Mar 19, 2014
RNAseqTRAPLINE	RNA sequencing data analysis in a Transparent Reproducible and Automated Pipeline - TRAPLINE.	mwolfien	★★★★★	rna-seq fastq tophat2 protein_interaction mirna_target_prediction	May 19, 2016
mt analysis 0.01 strand-specific (fastq single)		aun1	★★★★★		Mar 24, 2011
galaxy101-2015		aun1	★★★★★		Oct 28, 2015
metagenomic analysis	Generic workflow for performing a metagenomic analysis on NGS data.	aun1	★★★★★	metagenomics galaxy ngs	Aug 20, 2014
DEG:Tuxedo2 - 2 Conditions 1 Replicate each	Tuxedo DEG pipeline to compare 2 conditions	setempler	★★★★★		Oct 09, 2015

Obtain public workflows from <https://usegalaxy.org/> developed by the community.

or

Share own workflows within the group or IPK.

By default Galaxy assign 'fastq' data type to uploaded FASTQ files.

In this case the offset is not specified, and many tools do not recognize the data.

- *fastqillumina*: old illumina quality score encoding (offset 64, Illumina1.3+)
- *fastqsanger*: new Illumina1.8+ / Sanger quality score encoding
⇒ Adjust format during upload

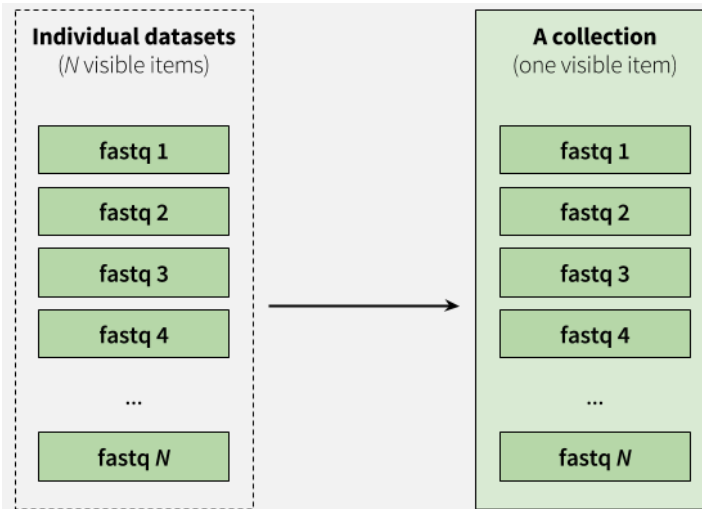
You added 1 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Type	Genome	Settings	Status
 input	1.2 KB	Auto-det...  	unspecified (?) 		 0%

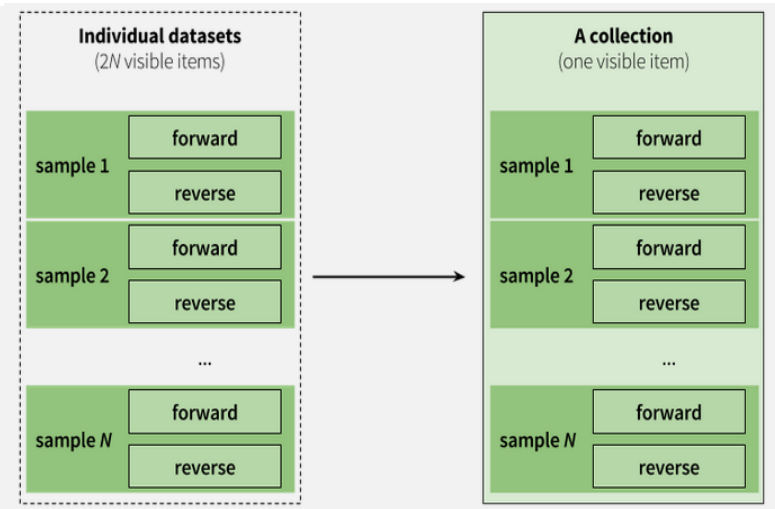
⇒ Change data type in your history

Collections allow users to handle and process a large number of samples at once.

Lists can contain an arbitrary number of elements



Paired list, containing a forward and reverse element each



<https://galaxyproject.org>

Uploading data into collections directly to bypass the need to upload single datasets into history first.

Collections allow users to handle and process a large number of samples at once.

1. Load your FASTQ samples in your history

2. Click „Operations on multiple datasets“ opens additional options in your history

3. Mark all files from your history that should be included in the list by clicking the small box in front of each sample. Go on with clicking „For all selected...“

4. This will open the following dialog:

5. Depending on the nature of sequencing data (paired or unpaired), two options are available:

- Hide datasets
- Unhide datasets
- Delete datasets
- Undelete datasets
- Build Dataset List
- Build Dataset Pair
- Build List of Dataset Pairs

4 NGS data

Paired dataset collections

Note: If your pairs are not named by „_1“ and „_2“ ending, you will probably see this message:

This should be corrected by typing the naming pattern of your data.
In this example „_R1“ and „_R2“:

Check the correct pairing of the files. Then click „Auto-Pair“.

The files are paired according to their name scheme. Name and create the list now.

Could not automatically create any pairs from the given dataset names. You may want to choose or enter different filters and try auto-pairing again. Close this message using the X on the right to view more help.

0 unpaired forward - (4 filtered out)

Choose filters Clear filters

1

Auto-pair

0 unpaired reverse - (4 filtered out)

2

(no datasets were found matching the current filters)

2 unpaired forward - (2 filtered out)

Choose filters Clear filters

R1

Auto-pair

2 unpaired reverse - (2 filtered out)

R2

CS1_R1.mini.trim.fq

Pair these datasets

CS1_R2.mini.trim.fq

ETC1_R1.mini.trim.fq

Pair these datasets

ETC1_R2.mini.trim.fq

2 pairs created: all datasets have been successfully paired

0 unpaired forward - (0 filtered out)

Choose filters Clear filters

R1

...

0 unpaired reverse - (0 filtered out)

R2

2 paired Unpair all

CS1_R1.mini.trim.fq → CS1.mini.trim ← CS1_R2.mini.trim.fq




ETC1_R1.mini.trim.fq → ETC1.mini.trim ← ETC1_R2.mini.trim.fq


Remove file extensions from pair names? ☒

Name: My paired list




Cancel

Create list

History   

search datasets 

GBS datasets
5 shown, 5 [deleted](#)

150.7 MB   

All None For all selected...

☐ **10: My paired list**
a list of paired datasets

☒ **4: CS1 R2.mini.trim.fq**

☒ **3: CS1 R1.mini.trim.fq**

☒ **2: ETC1 R1.mini.trim.fq**


☒ **1: ETC1 R2.mini.trim.fq**



New dataset collection:



- Click to view containing samples

[← Back to 2er original](#)

Select on collection 6
a list with 2 items

Add tags 








ETC1 R1.mini.trim.fq  

CS1 R1.mini.trim.fq  

291: featureCounts
on data 286

286: HISAT2 on
data 274 and data
273: aligned reads (BAM)

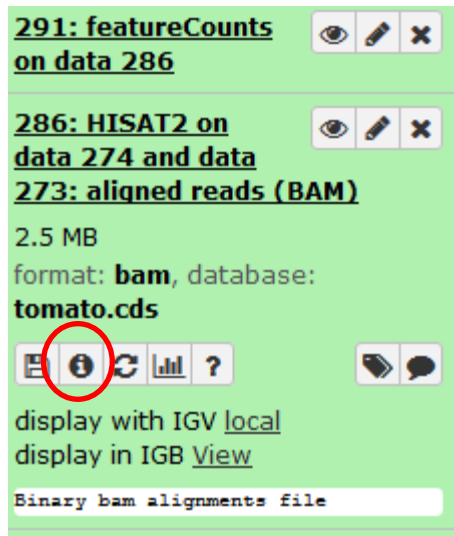
2.5 MB
format: **bam**, database:
tomato.cds

display with IGV [local](#)
display in IGB [View](#)

Binary bam alignments file

Example: HISAT mapping - where can I find mapping statistics?



Example: HISAT mapping - where can I find the mapping statistics?

=> Use samtools flagstat

=> Check for additional output!

Example: HISAT mapping - where can I find mapping statistics?

[291: featureCounts on data 286](#)

[286: HISAT2 on data 274 and data 273: aligned reads \(BAM\)](#)

2.5 MB
format: bam, d
tomato.cds

display with IGB
display in IGB

Binary bam align

HISAT

Dataset Information

Number:	286
Name:	HISAT2 on data 274 and data 273: aligned reads (BAM)
Created:	Mon 18 Feb 2019 01:36:49 PM (UTC)
Filesize:	2.5 MB
Dbkey:	tomato.cds
Format:	bam

Job Information

Galaxy Tool ID:	toolshed.g2.bx.psu.edu/repos/iuc/hisat2/hisat2/2.1.0
Galaxy Tool Version:	2.1.0
Tool Version:	/opt/galaxy/dependency_dir/_conda/envs/mulled-v1-3b104c294f65450b09ba89d24826c61eddd553d789c9f1ae48a29715de1b7 version 2.1.0 64-bit Built on login-node03 Wed Jun 7 15:53:42 EDT 2017 Compiler: gcc version 4.8.2 (GCC) Options: -O3 -m64 -r -DPOPCNT_CAPABILITY Sizeof {int, long, long long, void*, size_t, off_t}: {4, 8, 8, 8, 8, 8}

Tool Standard Output:	stdout
Tool Standard Error:	stderr

Tool Exit Code: 0



Empty, the content is redirected to your BAM file

Example: HISAT mapping - where can I find the mapping statistics?

[291: featureCounts on data 286](#)

[286: HISAT2 on data 274 and data 273: aligned reads \(BAM\)](#)

2.5 MB

format: bam, d

tomato.cds

display with IGB

display in IGB

Binary bam align

HISAT

Dataset Information

Number:	286
Name:	HISAT2 on data 274 and data 273: aligned reads (BAM)
Created:	Mon 18 Feb 2019 01:36:49 PM (UTC)
Filesize:	2.5 MB
Dbkey:	tomato.cds
Format:	bam

Job Information

Galaxy Tool ID: toolshed.g2.bx.psu.edu/repos/iuc/hisat2/hisat2/2.1.0

Galaxy Tool Version: 2.1.0

Tool Version: /opt/galaxy/dependency_dir/conda/bin/hisat2 version 2.1.0 64-bit Built on login node

Tool Standard Output: [stdout](#)

Tool Standard Error: [stderr](#)

Tool Exit Code: 0

```
19941 reads; of these:
 19941 (100.00%) were paired; of these:
   977 (4.90%) aligned concordantly 0 times
 18741 (93.98%) aligned concordantly exactly 1 time
   223 (1.12%) aligned concordantly >1 times
----
  977 pairs aligned concordantly 0 times; of these:
    118 (12.08%) aligned discordantly 1 time
----
 859 pairs aligned 0 times concordantly or discordantly; of these:
 1718 mates make up the pairs; of these:
 1253 (72.93%) aligned 0 times
  421 (24.51%) aligned exactly 1 time
   44 (2.56%) aligned >1 times
96.86% overall alignment rate
```

Special thanks to

- useGalaxy.eu team (especially Björn Grüning and Helena Rasche)
- BIT team
- IPK PostDoc board

